National Institute of Environmental Health Sciences

Biostatistics short course- II
Testing statistical hypotheses

Min Shi
Biostatistics Branch

Slides courtesy of Dr. Shyamal D. Peddada

Outline

- Basics of hypothesis testing.
  - Formulation of statistical hypotheses.
  - P-value & level of significance of a test.
  - Power and sample size calculations.

- Comparison of two or more groups.
  - Parametric procedures.
  - Nonparametric procedures.
  - Resampling based methods: Permutation and Bootstrap procedures.

- Multiple testing in high dimensional data.
  - Types of error rates.
  - Two important procedures for high dimensional data.

Basics of hypothesis testing

Formulation of hypotheses

- Null hypothesis H_0
  - Hypothesis of no difference or no association.
- Alternative hypothesis H_a
  - Hypothesis the researcher is trying to prove.

The basic idea

Assuming that the null hypothesis is true, is there "sufficient" evidence in the data to reject the null hypothesis in favor of the alternative hypothesis?

Example:

- Defendant is guilty (test substance is toxic).
- Defendant is innocent (test substance is not toxic).

Which is the null and which is the alternative hypothesis from a jury's (toxicologist's) point of view?

The basic idea

Null hypothesis: Defendant is innocent.

(No difference between the test substance and the vehicle control).

Alternative hypothesis: Defendant is guilty.

(There is difference between the test substance and the vehicle control).

The basic idea

Assuming that the defendant is innocent (there is no difference between the test substance and the vehicle control), how likely are we to see the following?

- Defendant's finger prints at the crime scene (half the treated animals developed tumors).

- An eye witness who saw the defendant at the crime scene (none of the control animals developed tumors).

Etc.

Caution

The following formulation of hypotheses is not valid.

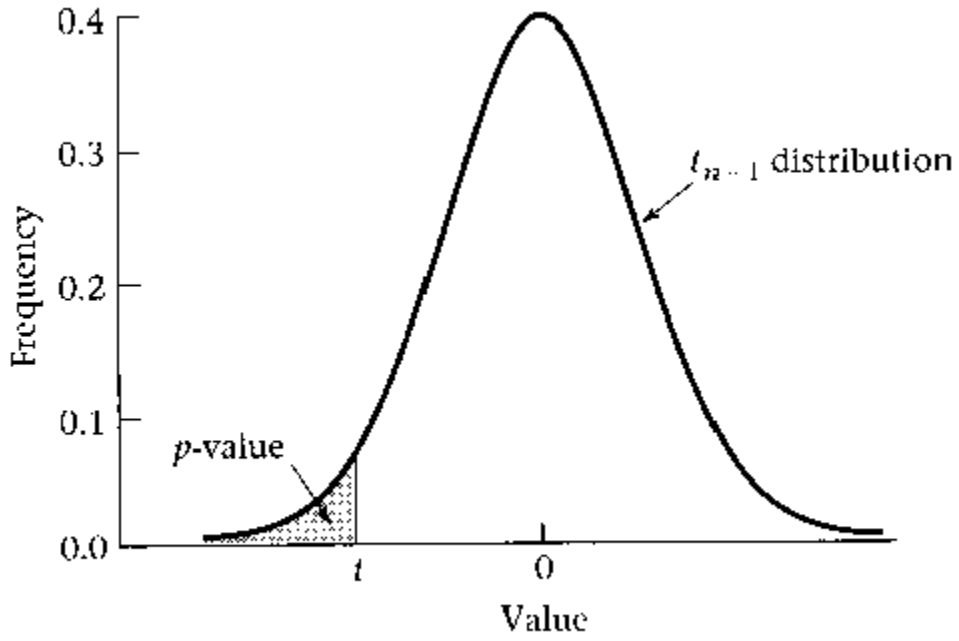
H_0 : There is a higher tumor incidence in the test chemical group than in the vehicle control group.

H_a : There is no difference in tumor incidence between the test chemical and vehicle control.

The p-value

Calculate the probability of outcomes as extreme as the ones observed – assuming the null hypothesis is true.

This is known as the p-value!

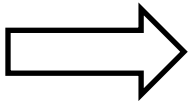


The p-value

Calculate the probability of outcomes as extreme as the ones observed – assuming the null hypothesis is true.

This is known as the p-value!

If p-value is small then it implies that you observed an event that is very unlikely to occur had the null hypothesis been true!



Reject the null hypothesis if the p-value is "small". The defendant is guilty (test substance is toxic)!

Level of significance

- How "small" is small? It is subjective!
  - This threshold is known as the "level of significance" against the null hypothesis. Denoted by α
  - Typically $\alpha = 0.05$
  - Thus you may reject the null hypothesis if the p-value < α

Interpretation

What should be the conclusion when the p-value is "large"?

Should it be "Accept the null hypothesis"? (i.e. defendant innocent).

No! Wrong conclusion.

Correct conclusion: We conclude that there is not sufficient evidence in the data to reject the null hypothesis at the specified level of significance! (fail to reject the null that the defendant is innocent)

Caution

In standard statistical framework you can never prove the null hypothesis since you are performing your statistical test assuming the null hypothesis is true!

You need to state the hypothesis you desire to prove as the alternative hypothesis. If the evidence is sufficiently strong against the null hypothesis you can reject the null in favor of the alternative hypothesis.

Bioequivalence Test

What if the hypothesis of interest is the equivalence of means of two populations?

e.g., Test the hypothesis that the mean expression of a gene is same between normal and tumor tissue

We can formulate the hypothesis as follows:

H_0: Mean expressions are not equivalent

$$\mu_1 - \mu_2 \leq -\delta \; or \; H_0: \; \mu_1 - \mu_2 \geq +\delta$$

H_a: Mean expressions are equivalent

$$-\delta < \mu_1 - \mu_2 < +\delta$$

where $[-\delta \, , +\delta]$is the allowable range for equivalence

Bioequivalence Test

Restate the hypotheses is as the following pair of hypotheses and test each of them.:

$$H_{01}: \mu_1 - \mu_2 \leq -\delta$$
$$H_{a1}: \mu_1 - \mu_2 > -\delta$$

and

$$H_{02}: \mu_1 - \mu_2 \geq +\delta$$
$$H_{a2}: \mu_1 - \mu_2 < +\delta$$

Conclude equivalence if and only if both H_{01} and H_{02} were rejected.

Formulation of alternative hypothesis

Two common types of alternative hypothesis.

Two-sided hypothesis

H_a : Mean of the treatment group is different from the that of the vehicle control.

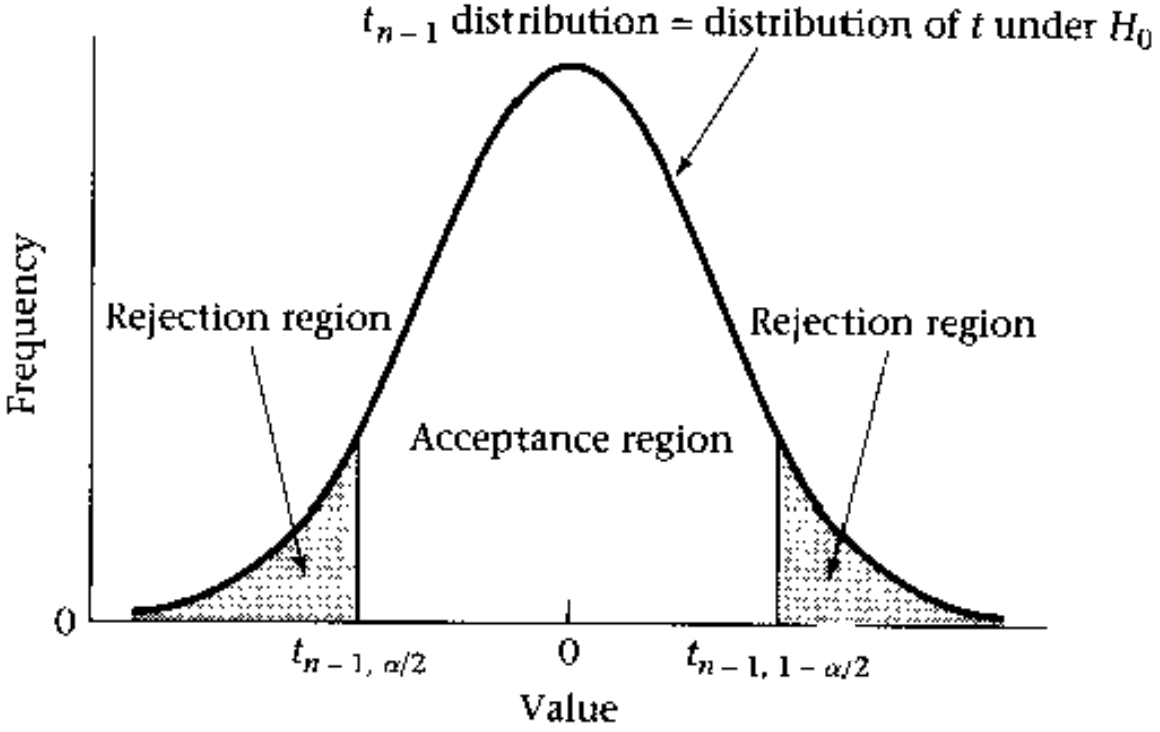
One-sided hypothesis

H_a : Mean of the treatment group is larger than that of the vehicle control.

More precisely …

$$H_0 : \mu_1 = \mu_2$$
$$H_a : \mu_1 \neq \mu_2 \quad \text{(Two-sided)}$$

Acceptance and rejection regions

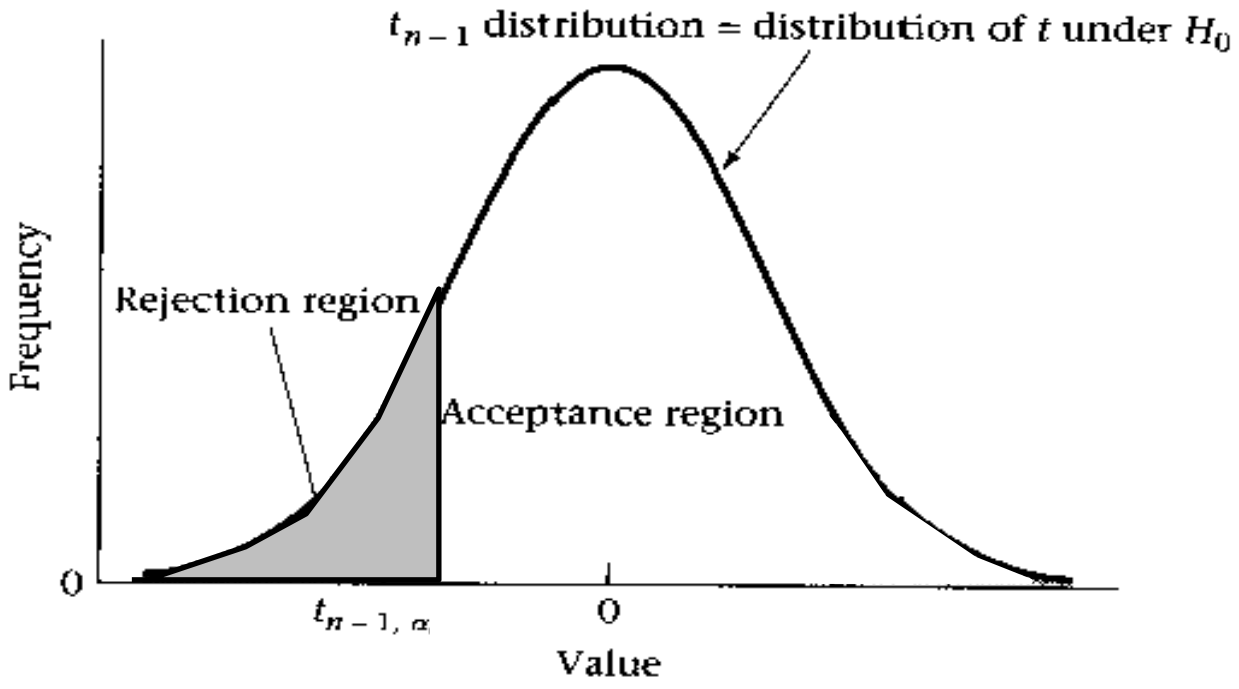


More precisely …

$$H_0 : \mu_1 = \mu_2$$

$$H_a : \mu_1 < \mu_2 \quad \text{(One-sided)}$$

Acceptance and rejection regions



Types of errors

All inferences are based on a sample of data (or evidence).

Consequently we are likely to make errors in our inferences.

Types of errors

- False positive = Falsely rejecting the null hypothesis (Type I error).
  - Declaring the defendant to be guilty when the person is innocent.
  - Declaring a chemical to be toxic when it is not really different from the vehicle control.
- False negative = Failing to reject the null when the alternative hypothesis is true (Type II error).
  - Failing to declare a guilty person to be guilty.
  - Failing to declare a toxic chemical to be toxic.

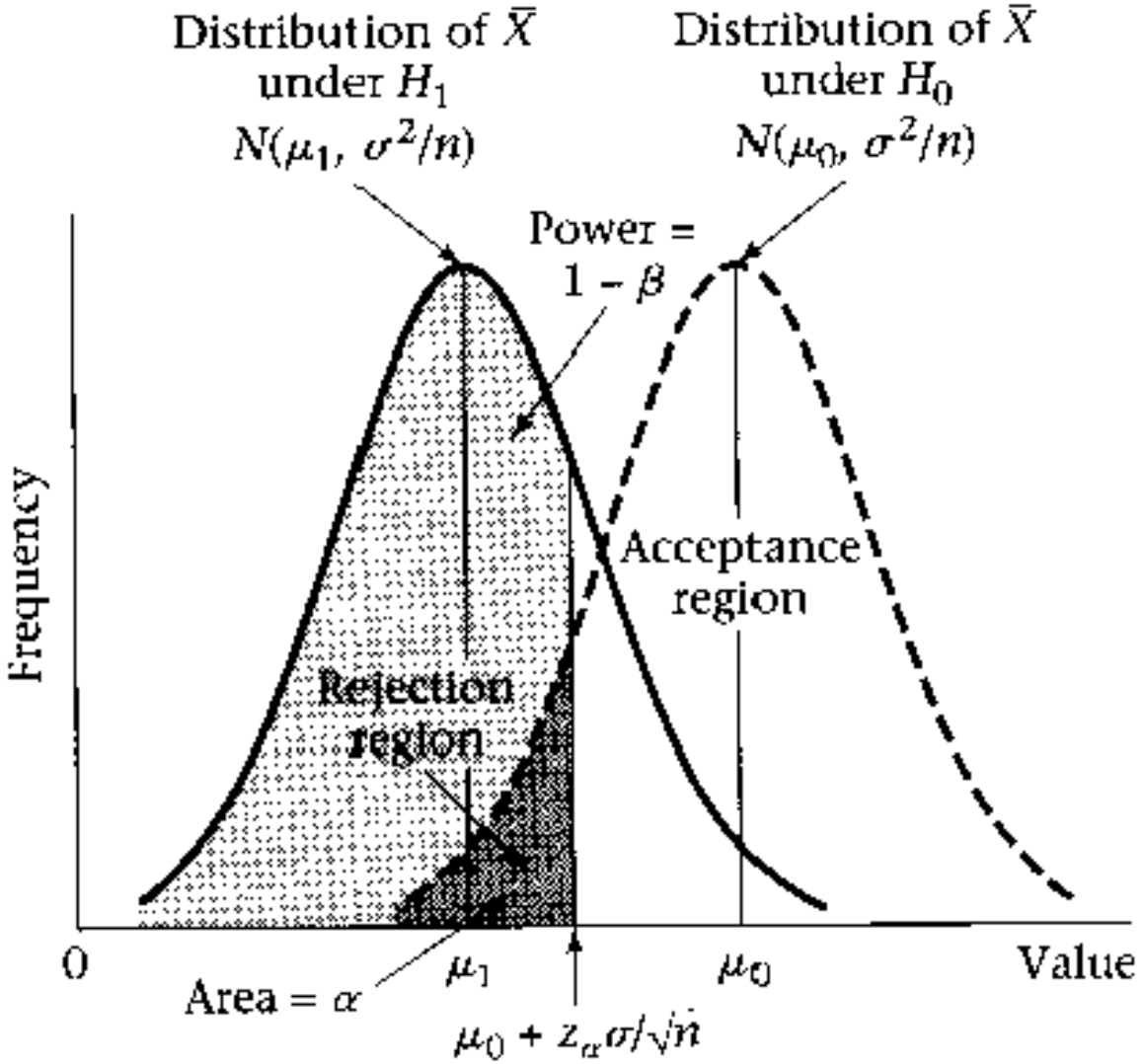
Types of errors

		State of Reality	
		H_0 True	H_0 False
Decision Made Based on Sample Data According to Rule	Do not reject H_0	Correct Decision	Error in Decision (Type II error)
	Reject H_0	Error in Decision (Type I error)	Power Pr(reject H_0|H_1)

Illustration of Power

$\mu = \mu_1 < \mu_0$

Power=
Pr(reject $H_0|H_1$)



Designing studies

- In addition to the underlying science, several factors play an important role in conducting good experiments/studies.
  - Randomization
    - To avoid any systematic bias.
  - Replication (loosely speaking "sample size")
    - To get a good estimate of the underlying variability in the data.
    - Impacts the power of a study.

Power and sample size

$$n = \frac{\sigma^2 (z_{1-\beta} + z_{1-\alpha/2})^2}{(\mu_0 - \mu_1)^2}$$

Power: Probability of rejecting the null hypothesis when the alternative hypothesis is true.

Some factors that impact sample size required in an experiment:

1. Power: n ↑ with increase in the desired power.
2. Level of significance: n ↓ with increase in the desired the level of significance.
3. Variability in the data: n ↑ with increase in variability in the data.

Power and sample size

$$n = \frac{\sigma^2 (z_{1-\beta} + z_{1-\alpha/2})^2}{(\mu_0 - \mu_1)^2}$$

4. Effect size: n ↓ with increase in the desired effect size.

5. Experimental design.

6. Type of alternative hypothesis (Very important when designing your study!).

7. Choice of the statistical test.

Factors influencing sample size calculations – An illustration

Sample size calculation for a power of 80% in a two-sample t-test with a 5% level of significance.

		Sample size	
Coefficient of variation (CV)	Effect size	One sided alternative	Two sided alternative
0.5	1.5	28	34
0.5	2	8	12
1	1.5	102	128
1	2	28	34
1.5	1.5	224	286
1.5	2	58	74
1.5	3	16	20

CV= Standard deviation/Mean

Check list for sample size calculation

- Are you interested in one-sided or two-sided hypothesis?
  - In case you want to compare more than 2 groups – talk to a Biostatistician!
- Determine the effect size you are hoping to detect. Choose a reasonable/realistic effect size.
  - Too small would result in a very large sample size!
  - Too large may result in an unrealistically small sample size and you actually may not have good power for a reasonable effect size.

Check list for sample size calculation

- Obtain an estimate of the anticipated variability in the data.
  - Always challenging since you have not done the experiment yet!

  Some tips you can use:

  1. Pilot studies conducted in the past.
  2. Similar experiments/studies published in the literature.
  3. Guess a plausible range of the data and divide by 4. This might provide an estimate under the assumption that the data are likely to be normally distributed.

Comparison of two or more groups…

Three classes of methods

A. Parametric methods.
   - Underlying probability distribution is known (e.g. normally distributed data).

B. Nonparametric methods (Distribution free methods).
   - Underlying probability distribution is not known.

C. Re-sampling based methods (Useful for bioinformatics).
   - Underlying probability distribution is not necessarily known.
   - Some of the resampling methods are more flexible than the standard nonparametric methods.

Parametric methods

Example 1

- Consider an assay resulting in the following data:

Replicate	Control	Treatment
1	2.73	2.04
2	1.57	7.39
3	1.02	4.37
4	1.6	6.53
5	0.41	
6	3.44	

- Question: Is the mean of the treatment group significantly different from the mean of the control?

"Typical" strategy – Parametric procedure

- Compute:
  - the sample means of the two groups.
  - the sample standard deviations.

Group	Mean $\bar{x}$	Standard Deviation S
Control	1.80	1.11
Treatment	5.08	2.39

- **"Pooled t-test"** statistic:

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{S_{pooled}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$S_{pooled} = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}$$

- P-value

Back to Example 1

- P-value = **0.0176**

- Conclusion: The result is significant at 5% level of significance. Thus the mean of treatment group is significantly different from the mean of the control group at 5% level of significance.

- Question: Is this conclusion valid?

When is the pooled t-test valid?

Assumptions made in pooled t-test

1*. Samples within each group are a simple random sample.

- Avoids any systematic bias in your samples.
- Observations are independent.

2*. Samples between groups are independent.

- e.g. not repeated measures.

Assumptions made in pooled t-test

3*. Data within each group are approximately normally distributed!

- Not a critical assumption if the sample sizes are "large" or if the distributions are approximately symmetric.

4*. The population variance of the two groups is same!

$$\sigma_1^2 = \sigma_2^2$$

Known as homoscedasticity (or variances are homogeneous).

Questions

- How do we detect if $\sigma_1^2 \neq \sigma_2^2$?

  Known as heteroscedasticity (or heterogeneous) variances.

- If we conclude heteroscedasticity then how do we compare the means?

Methods for detecting heteroscedasticity

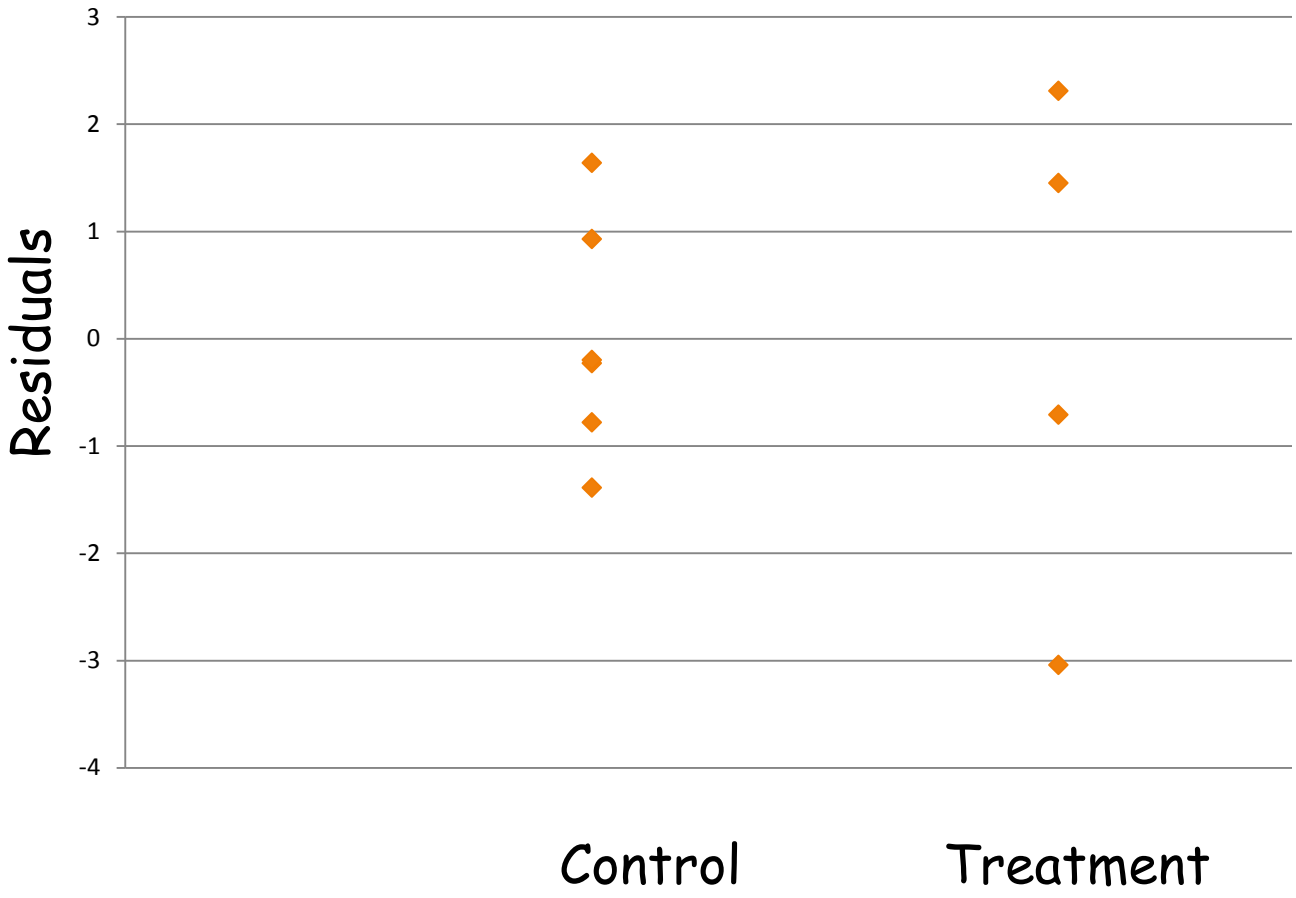
1. Formal statistical test (an F-test).
   - The data need to be normally distributed.
2. Graphical method using residuals.

Replicate	Control data	Control residual	Trt. data	Trt. residual
1	2.73	2.73 - 1.80	2.04	2.04 - 5.08
2	1.57	1.57 - 1.80	7.39	7.39 - 5.08
3	1.02	1.02 - 1.80	4.37	4.37 - 5.08
4	1.6	1.60 - 1.80	6.53	6.53 - 5.08
5	0.41	0.41 - 1.80		
6	3.44	3.44 - 1.80		

Graphical method using residuals

Replicate	Control residual	Treatment residual
1	0.93	-3.04
2	-0.23	2.31
3	-0.78	-0.71
4	-0.20	1.45
5	-1.39	
6	1.64	

Residual plot to detect heteroscedasticity



Conclusion?

- Even though the two-sided P-value (**0.0176**) based on the pooled t-test appears to be significant, this conclusion may not be valid since the two groups appear to have different variances.

Some strategies when the
pooled t-test is not appropriate
(Known as the Behrens-Fisher problem)

Welch's and other approximations (available in some standard packages)

The t-statistic is different from the pooled t-test:

$$\frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

The degrees of freedom are calculated from the data and is NOT $n_1 + n_2 - 2.$

Back to the Example 1

- Compute:

"Unequal variance t-test" statistic:

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} = 2.57$$

Group	Mean	Standard Deviation
Control	1.80	1.11
Treatment	5.08	2.39

- Welch's P-value using EXCEL = **0.0640** - NOT significant at 5% level of significance!

Effect of unequal variance on the Pooled t-test

A. If the two groups have the same sample size then the pooled t-test should be okay in general.

B. If $\sigma_1^2 < \sigma_2^2$ and $n_1 > n_2$ then pooled t-test may have.

higher false positive rate than the desired nominal level.

- Thus you can't be too sure about the significant result observed in the data.

C. If $\sigma_1^2 < \sigma_2^2$ and $n_1 < n_2$ then pooled t-test may have.

smaller false positive rate than the desired nominal level.

- Hence may result in loss of power.

Question

Why not use the **unequal variance t-test** for all data sets since it requires one less assumption?

Answer: If the two populations have equal variances then the **unequal variance t-test** will be less powerful than the pooled t-test.

A simple alternative strategy

- Perform transformations to data and then apply pooled t-test.

1. Log-transformation –perhaps the data are log-normally distributed.

$$Y \to \log(Y + c)$$

2. More generally perform Box-Cox power transformations.

$$Y \to \frac{(Y + c)^d}{d}, d \neq 0$$

B. Nonparametric methodology

- Useful when distribution of the data within each group is unknown or non-normal!

Wilcoxon Rank Sum test

Control: 2.73, 1.57, 1.02, 1.60, 0.41,3.44

Treatment: 2.04, 7.39, 4.37, 6.53

Mix the two samples and rank from smallest to largest

- Combined data:

  0.41, 1.02, 1.57, 1.60, 2.04, 2.73, 3.44, 4.37, 6.53, 7.39

- Rank:

  1, 2, 3, 4, 5, 6, 7, 8, 9, 10

Rank data:

- Control: 1, 2, 3, 4, 6, 7
- Treatment: 5, 8, 9, 10

Wilcoxon Rank Sum test

Rank data:

- Control: 1, 2, 3, 4, 6, 7 Sum=23
- Treatment: 5, 8, 9, 10 Sum=32

If the sum of the Control group is "unusually" small or large, you reject the null hypothesis that the two groups have same distribution. The critical values are obtained by Wilcoxon rank sum test tables or normal approximation (if the sample size is large enough).

P-value is 0.0428

What assumptions are required for Wilcoxon Rank Sum test?

Assumptions made in pooled t-test

1*. Samples within each group are a simple random sample.

- Avoids any systematic bias in your samples.
- Observations are independent.

2*. Samples between groups are independent.

- e.g. not repeated measures.

Assumptions made in ~~pooled t test~~

- 1*. Samples within each group are a simple random sample.
  - Avoids any systematic bias in your samples.
  - Observations are independent.
- 2*. Samples between groups are independent.
  - e.g. not repeated measures.

Assumptions made in pooled t-test

3*. Data within each group are approximately normally distributed!

- Not a critical assumption if the sample sizes are "large" or if the distributions are approximately symmetric.

4*. The population variance of the two groups is same!

Assumptions made in ~~pooled t test~~

3*. Data within each group are approximately normally distributed!

Not needed!

– Not a critical assumption if the sample sizes are "large" or if the distributions are approximately symmetric.

4*. The population variance of the two groups is same!

Assumptions made in ~~pooled t test~~

3*. Data within each group are approximately normally distributed!

Not needed!

– Not a critical assumption if the sample sizes are "large" or if the distributions are approximately symmetric.

3* ~~4*~~. The population variance of the two groups is same!

Some comments

- Median test is another nonparametric test commonly used – compare the medians.
- Wilcoxon rank sum test is also the Mann-Whitney test. Some times these tests are also known as Mann-Whitney-Wilcoxon test.
- Wilcoxon rank sum test is also a special case of Kruskal-Wallis test.

C. Resampling methods

Two general schemes

- Methods that enable us to derive the distribution of a test statistic under the null hypothesis.
  - Permutation:
    - Exchange labels on data points: random draws without replacement from the mixed sample.
  - Bootstrap (numerous variations exist):
    - Bootstrap -1: Random draws with replacement from the mixed sample.
    - Bootstrap -2: Bootstrap the residuals.

Permutation test

1. Compute the test statistic **T** using the given data.
2. Combine the two sets of samples into one (n1+n2).
   - Assign a random sample of n1 observations (without replacement) to the Control group. The remaining n2 are assigned to the Treatment group.
3. Construct the test statistic using the above null data. Denote it by **T*. [center at the difference of the sample means]**
4. Repeat the above process a large number of times.
5. Determine the proportion of times **T* > |T|.** Multiply by 2 to get a p-value for the two-sided test.

Example: Permutation test

Ref: http://faculty.washington.edu/kenrice/sisg/SISG-08-06.pdf

Generate date: 10 exposed subjects and 10 unexposed subjects. 1) $\mu_0 = \mu_1 = 0$;
2) $\mu_0 = 0$ and $\mu_1 = 1$

```
> set.seed(11)
> exposure<-rep(c(0,1), c(10,10))
> null.y<-rnorm(20)
> alt.y<-rnorm(20, mean=exposure)
> null.diff<-mean(null.y[exposure==1])-
mean(null.y[exposure==0])
> alt.diff<-mean(alt.y[exposure==1])-mean(alt.y[exposure==0])
> null.diff [1]
   -0.2390708
 > alt.diff [1]
   0.9939836
```

Example: Permutation test

Generate permutations:

```
one.test <- function(x,y) {
  xstar<-sample(x)
  mean(y[xstar==1])-mean(y[xstar==0]) }

many.truenull <- replicate(1000, one.test(exposure, null.y))
many.falsenull <- replicate(1000, one.test(exposure, alt.y))
```

sample(x): generate random permutation

```
> exposure
[1] 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1
> sample(exposure)
[1] 0 0 1 0 1 0 0 0 1 1 0 1 0 1 1 1 1 0 0 1
```

Example: Permutation test

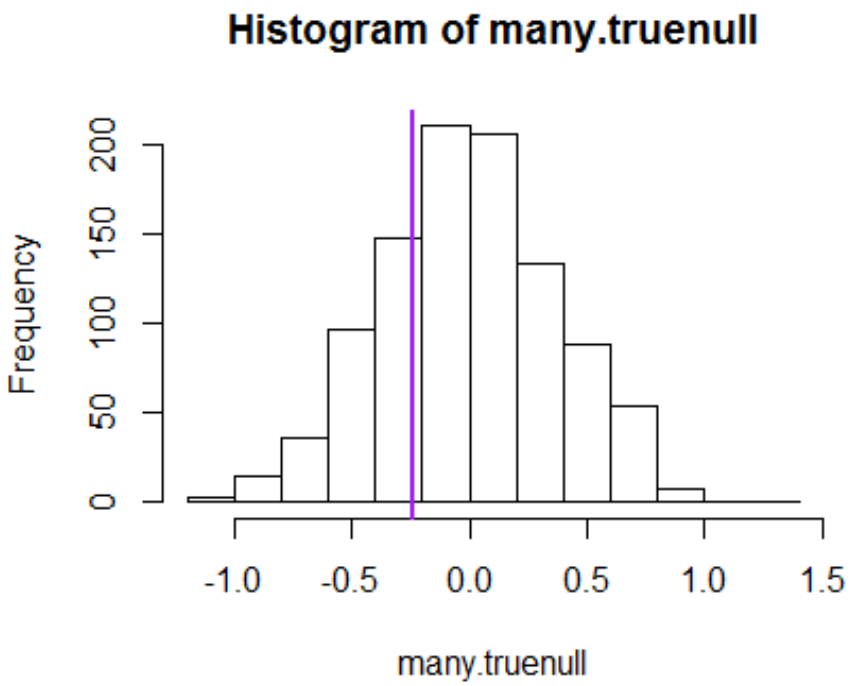
R codes for plotting

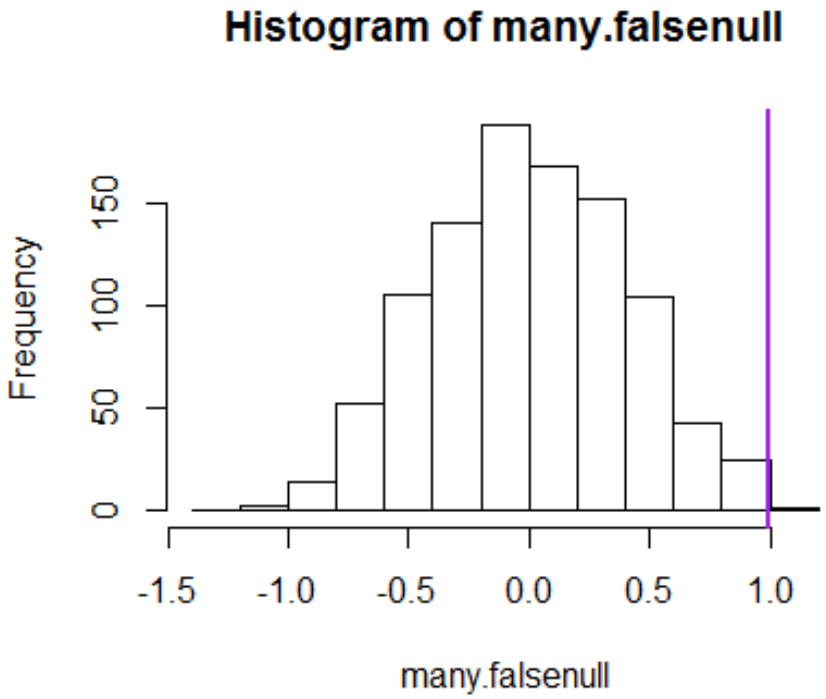
```
> hist(many.truenull)
> abline(v=null.diff, lwd=2, col="purple")
> mean(abs(many.truenull) > abs(null.diff))
[1] 0.52
> hist(many.falsenull)
> abline(v=alt.diff, lwd=2, col="purple")
> mean(abs(many.falsenull)
> abs(alt.diff))
[1] 0.006
```

Example: Permutation test

$\mu_0 = \mu_1 = 0$

$\mu_0 = 0$ and $\mu_1 = 1$





Percent of times |permuted mean| > |observed mean|:

0.52

0.006

Bootstrap -1

1. Compute the test statistic **T** using the given data.
2. Combine the two sets of samples into one (n1+n2).

   - Assign a random sample of n1 observations (with replacement) to the Control group and a random sample of n2 (with replacement) to the Treatment group.

3. Construct the test statistic using the above null data. Denote it by **T*. [center at the difference of the sample means]**

4. Repeat the above process a large number of times.

5. Determine the proportion of times **T* > |T|.** Multiply by 2 to get a p-value for the two-sided test.

Comments on Permutation and Bootstrap -1

- If the sample sizes are large then the p-values from the two methods are approximately same.
- Both methods make same assumptions as the Wilcoxon rank sum test does.
- Neither method is suitable if the two groups have different variances.

Back to example 1

The bootstrap p-value using 10,000 bootstrap samples = 0.0610.

Conclusion: Do not reject the null that the two groups have same mean at 5% level of significance.

More than two treatment groups… ANOVA
Interval scale data

Three classes of methods

A. Parametric methods.

B. Nonparametric methods (Distribution free methods).

C. Re-sampling based methods.

Parametric methods

The global/complete null hypothesis

$H_0 : \mu_1 = \mu_2 = ... = \mu_p$

(All group means are equal)

H_a : At least one group mean is different from others.

Example 2

- Consider an assay resulting in the following data:

Replicate	Control	Dose 1	Dose 2	Dose 3
1	10.919	10.936	11.636	2.756
2	10.880	8.933	12.051	8.140
3	10.412	10.691	11.302	8.987
4	10.838	9.392	14.009	11.523
5	10.491	9.349	12.771	2.740
6	9.109	8.869	14.239	11.372

The Analysis of Variance (ANOVA) – the basic idea

- Compute the means of the data from each group.

	Control	Dose 1	Dose 2	Dose 3
Sample Mean	10.441	9.695	12.668	7.586

- Basic idea of F-test:
  - If sample means of the four groups are "far apart" relative to the within group variation then you may want to reject the null hypothesis.

Analysis of Variance – More precisely

- The F-test used in ANOVA:

$$\frac{\text{Mean variability between groups}}{\text{Mean variability within groups}}$$

- The numerator degrees of freedom = number of groups -1
- The denominator degrees of freedom = total sample size - number of groups

Back to Example 2

- Means and SD for the data are:

	Control	Dose 1	Dose 2	Dose 3
Mean	10.441	9.695	12.668	7.586
Standard Deviation	0.687	0.895	1.232	3.973

- The P-value from F-test = **0.0055.**
- Is this p-value valid? Verify the assumptions similar to those described for the t-test.

What can I infer?

- The mean of at least one of the four groups is significantly different from the others at level of significance of 0.01 because the p-value is **0.0055**.

- Can anything be said about which mean is different or how the means differ from each other?

Post-hoc Analysis

A researcher is often interested in:

- Pairwise comparisons of treatment groups vs. control group.
- All pairwise comparisons.
- Trend test to detect dose-response.

Post-hoc analysis can be used

Post-hoc multiple comparisons

Table 2. Features of the most commonly used post-hoc tests (modified from Abacus Concepts 1993 and Armstrong *et al.*, 2000)

Method	Equal N F	Normality	Use	Error control	Protection
Fisher PLSD	Yes	Yes	Yes	All	Most sensitive to Type 1
Tukey-Kramer HSD	No	Yes	Yes	All	Less sensitive to Type 1 than Fisher PLSD
Spjotvoll-Stoline	No	Yes	Yes	All	As Tukey-Kramer
Student-Newman Keuls (SNK)	Yes	Yes	Yes	All	Sensitive to Type 2
Tukey-Compromise	No	Yes	Yes	All	Average of Tukey and SNK
Duncan's Multiple Range	No	Yes	Yes	All	More sensitive to Type 1 than SNK
Scheffé's S	Yes	No	No	All	Most conservative
Games/Howell	Yes	No	No	All	More conservative than majority
Dunnett's test	No	No	No	T/C	More conservative than majority
Bonferroni	No	Yes	Yes	All, TC	Conservative

Abbreviations: PLSD = Protected least significant difference, HSD = Honestly significant difference.
T = treatment groups, C = Control group, Column 2 indicates whether equal numbers of replicates (N) in each treatment group are required or whether the method can be applied to cases with unequal 'N'. Column 3 indicates whether a significant between treatments F ratio is required before post-hoc tests can be applied and columns 4 and 5 whether the method assumes equal variances in the different treatments and normality of errors respectively. The final column indicates the relative degree of protection against type 1 and type 2 errors.

http://eprints.aston.ac.uk/9317/1/Statnote_6.pdf

Formulation of statistical hypothesis!

When a researcher knows what he/she wants to compare the F-test may not be the most appropriate test.

Alternative more powerful methods of analysis are available depending upon the scientific question of interest.

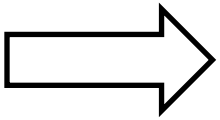
An example of power comparisons in a dose-response study (a simulated data)

- Level of significance = 0.05
- Number of dose groups = 4
- Sample size at each dose = 6
- Standard deviation of each group = 1

Mean patterns	Power using F-test	Power using trend test
(1, 1.25, 1.5,1.75)	0.157	0.355
(1, 2, 2, 2)	0.335	0.547
(1, 2, 2.5, 3)	0.795	0.958

- Substantial gains in power by testing for trend rather than using the standard ANOVA based F-test.

Formulation of statistical hypothesis!



Formulation of the statistical hypothesis is a very important step before performing any analysis of data.

Multiple testing in genomics and other high dimensional data analysis

Discoveries in microarray data and other high dimensional data

- Consider a microarray experiment with the following experimental design:
  - Control group - 6 animals
  - Treatment group - 6 animals
  - Affy chip consisting of m = 45000 probes.

Questions: Identify differentially expressed probes.

- Identify probes that are significantly up regulated in the treatment group.
- Identify probes that are significantly down regulated in the treatment group.

Discoveries in high dimensional data

- **Problem of interest**:
  - Compare two (or more) groups on the basis of m = 45000 probes! Thus 45000 statistical tests are being performed!
- **False positive rate accumulates**

Number of probes (m)	Probability falsely rejecting at least one null hypothesis (assuming probes are independent)
1	0.05
2	0.10
3	0.14
5	0.23
45000	1

Two commonly used error rates

- Family Wise Error Rate (FWER).
  - Probability of falsely rejecting at least one null hypothesis among all hypotheses tested.
- False Discovery Rate (FDR).
  - The expected proportion of false discoveries among all discoveries made.

Classification of m hypothesis tests

	Null hypothesis is True	Alternative hypothesis is True	Total
Declared significant	***V***	***S***	***R***
Declared non-significant	***U***	***T***	***m-R***
Total	m_0	$m-m_0$	***m***

- FWER = Pr(V>=1) or equivalently, FWER = 1- Pr(V = 0)
- FDR = E (V/R) = E(V/(V + S))

Control of FWER …

Controlling FWER

Bonferroni method

Suppose the microarray consists of "m" probes on the chip.

1. Compute the standard p-value for each probe.

2. Multiply the p-value by "m". If the result is more than 1 then set it to 1. This is called the Bonferroni adjusted p-value.

Decision rule:

For a given probe if its Bonferroni adjusted p-value is less than 0.05 then you conclude that it is differentially expressed at FWER of 0.05.

This procedure **can be applied very broadly**, but is **conservative.**

Control of FDR …

Controlling FDR

Benjamini-Hochberg procedure (Step-up procedure)

Suppose there are "m" probes on the microarray.

1. Compute p-value for each probe.
2. Sort probes by their p-values. $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(m)}$
3. Let $\alpha_{(i)} = \frac{i\alpha}{m}$
4. Identify the largest index r such that

$$p_{(r)} \leq \alpha_{(r)} \text{ and } p_{(r+1)} > \alpha_{(r+1)}$$

BH procedure

Reject all null hypotheses $H_{(1)}, H_{(2)}, \ldots, H_{(r)}$ corresponding to the p-values $p_{(1)}, p_{(2)}, \ldots, p_{(r)}$

An illustration

- Suppose we have a microarray containing 15 probes.
- The p-values based on standard t-test are given in the table.

Probe	P-value
Probe1	0.0043
Probe2	0.0037
Probe3	0.0008
Probe4	0.0001
Probe5	0.0042
Probe6	0.3098
Probe7	0.1112
Probe8	0.1712
Probe9	0.9676
Probe10	0.715
Probe11	0.0216
Probe12	0.5526
Probe13	0.7577
Probe14	0.023
Probe15	0.1545

Probe	Sorted P-value	Bonferroni p-value	Benjamini-Hochberg threshold
Probe4	0.0001	0.0015	0.003333
Probe3	0.0008	0.012	0.006667
Probe2	0.0037	0.0555	0.01
Probe5	0.0042	0.063	0.013333
Probe1	0.0043	0.0645	0.016667
Probe11	0.0216	0.324	0.02
Probe14	0.023	0.345	0.023333
Probe7	0.1112	1	0.026667
Probe15	0.1545	1	0.03
Probe8	0.1712	1	0.033333
Probe6	0.3098	1	0.036667
Probe12	0.5526	1	0.04
Probe10	0.715	1	0.043333
Probe13	0.7577	1	0.046667
Probe9	0.9676	1	0.05

$$\alpha_{(i)} = \frac{i}{m}\alpha$$

Here m= 15

$\alpha = 0.05$

i=1, 2, …, 15

Conclusion

- FWER controlling methods (e.g. Bonferroni method) tend to be more conservative compared to FDR controlling methods (e.g. BH procedure).
  - i.e. fewer number of probes will be selected as significant if FWER is controlled instead of FDR.

Questions?

TESTING FOR THE EQUALITY OF TWO VARIANCES

$H_0 : \sigma_1^2 = \sigma_2^2$ versus $H_1 : \sigma_1^2 \neq \sigma_2^2$

F distribution Let $X \sim \chi^2_{d_1}$ and $Y \sim \chi^2_{d_2}$. Assuming X is independent of Y, then

$$\frac{X/d_1}{Y/d_2} \sim F_{d_1,d_2}$$

where F_{d_1,d_2} is called the F distribution with degrees of freedom d_1 and d_2.

Wilcoxon Rank Sum test- normal approximation method

R_1 The sum of the ranks in the first sample. Under H_0 (when there are no ties),

$$E(R_1) = \frac{n_1(n_1 + n_2 + 1)}{2}$$

$$Var(R_1) = \frac{n_1 n_2(n_1 + n_2 + 1)}{12}$$

(a) Compute

$$T = \left[\left|R_1 - \frac{n_1(n_1 + n_2 + 1)}{2}\right| - \frac{1}{2}\right] / \sqrt{\frac{n_1 n_2}{12}(n_1 + n_2 + 1)}$$

if there are no ties

If $T > z_{1-\alpha/2}$, then reject H_0. Otherwise, accept H_0.

- Calculation of T is more complicated if there are ties

Bootstrap – 2: Bootstrap residuals

Basic difference between this method and Bootstrap-1 method is in Step 3.

- Details omitted.

Comments on bootstrapping residuals

Strengths

1. Distribution free. The samples can be from any continuous distribution.

2. No need to verify if the variances are equal or not.

3. Since it may not be easy to verify if the variances are equal or not in a high dimensional data , it is the ideal method for analyzing large scale genomic data.

4. This method can be extended to more complex modeling situations.

Potential weakness

1. May have a smaller power than the pooled t-test if the data satisfy the assumptions 3* and 4* required by t-test.

Equation $$\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \sim N(0,1)$$

Two-sample t test for independent samples with unequal variances (Satterthwaite's method) Let

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}},$$

$$d' = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1 - 1) + (s_2^2/n_2)^2/(n_2 - 1)}$$

and d'' be the nearest integer to d'.

If $t > t_{d'',1-\alpha/2}$ or $t < -t_{d'',1-\alpha/2}$, then reject H_0;